UNIFIED GRAPH ANOMALY DETECTION WITH TARGET-CONTEXT EVIDENCE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In graph-structured data, anomalous entities at different granularity levels (i.e., nodes, edges, and graphs) are often inherently interconnected. Motivated by such dependencies, unified graph anomaly detection (GAD) has recently emerged as a new research direction, aiming to detect multi-level anomalies within a single model. Although existing unified GAD approaches integrate multi-level detection tasks into a single framework, they typically convert detection objects at different granularities into a fixed subgraph format, overlooking the granularity-specific information required for anomaly detection at different levels. Moreover, existing methods still rely on level-specific detection heads or task-dependent references to support anomaly detection at different granularities, limiting knowledge integration throughout the end-to-end detection pipeline and preventing full task unification. To address these limitations, we propose **Uni**fied **T**arget-context **R**eformulation **A**nd **C**ontextual **E**vidence **R**econstruction (**UniTRACER**) for unified GAD. To preserve granularity-specific information with a general task formulation, we introduce a target-context joint modeling paradigm and emphasize context importance based on the specific detection target. Moreover, to consolidate multi-level anomaly detection under a unified detection objective, we characterize the target-context relation and importance as shared detection evidence among levels, and then reframe unified GAD as an evidence reconstruction task. Together, these designs enable UniTRACER to retain granularity-specific anomaly cues while facilitating end-to-end knowledge sharing through a common task formulation and detection objective. Comprehensive experiments on 17 datasets demonstrate that UniTRACER significantly outperforms state-of-the-art granularity-specific and unified GAD methods.

1 INTRODUCTION

Graph anomaly detection (GAD) aims to identify abnormal samples that significantly deviate from the majority in graph-structured data Pan et al. (2025). By detecting anomalous entities at node, edge, or graph levels, GAD can safeguard a wide range of real-world applications, such as identifying malicious users Qiao & Pang (2023), money-laundering transactions Han et al. (2026), and toxic chemicals Wang et al. (2025). Following the conventional task-specific graph learning paradigm, existing GAD studies typically develop a granularity-specific detector that focuses on identifying anomalies at either the node Liu et al. (2021b), edge Liu et al. (2024), or graph level Li et al. (2023). Nevertheless, this isolated formulation overlooks potential cross-level dependencies that commonly arise in real-world scenarios. For example, the toxicity of a chemical molecule may be caused by a toxic functional group, while malicious bot accounts may connect with each other to form a fraudulent network Lin et al. (2024).

To exploit such cross-level dependencies, **unified GAD** has recently emerged as a new research direction, aiming to detect multi-level anomalies within a single model Lin et al. (2024); Song et al. (2025). UniGAD pioneers this direction by transforming anomalies at different levels into a unified subgraph representation, and integrates multi-level knowledge through a stitch network under supervised joint training Lin et al. (2024). UniForm further extends unified GAD to the unsupervised setting by adopting meta-learning as its training scheme Song et al. (2025). This multi-task learning paradigm boosts anomaly detection knowledge transfer across different granularities, safeguarding complex real-world graph applications with hierarchical anomalies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

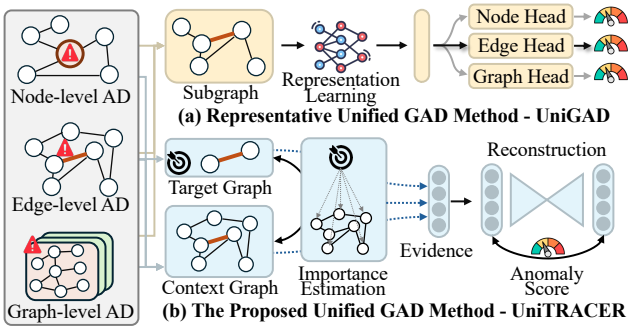


Figure 1: Comparison of unified GAD paradigms.

Although existing unified GAD approaches integrate multi-level detection tasks within a single framework, they unify detection objects at different granularities into a fixed format, neglecting the **granularity-specific information** required by different detection tasks. For example, node-level GAD typically relies on fine-grained local discrepancies between the target node and its neighborhood Qiao & Pang (2023), whereas graph-level GAD requires more global structural and contextual information to characterize anomalous patterns at larger scale Liu et al. (2023a). However, existing methods mainly homogenize node-, edge-, and graph-level detection objects into subgraphs with a fixed scheme Lin et al. (2024); Song et al. (2025), overlooking their distinct demands on the current detection objective Lin et al. (2024); Song et al. (2025). As a result, the model may struggle to distinguish task-relevant anomaly cues from irrelevant contextual information. Motivated by this limitation, we formulate a key research question for unified GAD method design: ***RQ1 - How can we capture granularity-specific information for anomaly detection within a unified task format?***

On the other hand, while existing methods combine multi-level anomaly detection within a single framework, node-, edge-, and graph-level anomaly detection are still treated as distinct detection tasks, rather than being formulated under a **unified detection objective**. Specifically, UniGAD retains separate prediction heads for different anomaly levels Lin et al. (2024), whereas UniFORM relies on task-dependent reference information to support detection at different granularities Song et al. (2025). Such a fragmented design hinders end-to-end knowledge integration throughout the entire detection pipeline and undermines true task unification. Moreover, this challenge is further amplified in the unsupervised setting, where anomaly labels are unavailable to explicitly coordinate detection objectives of different granularities. This naturally leads to our second research question: ***RQ2 - How can multi-level anomaly detection be formulated under a unified detection objective?***

Motivated by the above two research questions, we propose **Uni**fied **T**arget-context **R**eformulation **A**nd **C**ontextual **E**vidence **R**econstruction (**UniTRACER**) for unified GAD, with its sketch map shown in Figure 1. UniTRACER preserves granularity-specific information with a general task formulation, while consolidating multi-level anomaly detection under a unified detection objective. To address ***RQ1***, we introduce a *target-context joint modeling* paradigm that reformulates anomaly detection at different levels as assessing a target graph with the support of its context. To explicitly emphasize the information relevant to each detection target, UniTRACER estimates context importance conditioned on the current target as the core learning objective. In this way, node-, edge-, and graph-level anomalies share a common task formulation while retaining the granularity-specific information required for detection. Moreover, to handle ***RQ2***, we characterize the target-context relation and importance as shared detection evidence across anomaly levels, transforming their level-specific detection characteristics into a common evidence space. Built upon it, we formulate unified GAD as an *evidence reconstruction task* and employ a shared autoencoder to jointly learn normal patterns across levels. The resulting reconstruction error serves as a unified anomaly indicator, enabling multi-level anomalies to share a single detection objective. Extensive experiments across multiple datasets demonstrate the effectiveness of our method.

2 Preliminaries and Related Work

In this section, we present the preliminaries and briefly review the related work. A more comprehensive review of the related literature is provided in Appendix A.

Graph anomaly detection (GAD). GAD aims to identify abnormal objects in graph-structured data. As abnormality can manifest at node, edge, and graph levels, most existing approaches are granularity-specific to a particular detection level. **Node-level GAD**, as the most extensively investigated setting, is typically addressed through either unsupervised objectives, such as reconstruc-

tion Ding et al. (2019), contrastive learning Liu et al. (2021b); Pan et al. (2023), affinity modeling Qiao & Pang (2023), and denoising He et al. (2024); Li et al. (2025), or supervised binary classification with anomaly-sensitive network architectures Tang et al. (2022). **Edge-level GAD** focuses on detecting abnormal links by modeling structural dependencies and interaction patterns between nodes Duan et al. (2020); Liu et al. (2023c); Zheng et al. (2019); Liu et al. (2023b). **Graph-level GAD** identifies anomalous graphs from a graph collection and explores various learning paradigms, including transformation Qiu et al. (2022), reconstruction Kim et al. (2024), distribution modeling Cao et al. (2024), prototype learning Yang et al. (2025); Xu et al. (2026), and structural representation learning Li et al. (2023); Kim et al. (2024); Fu et al. (2026). Nevertheless, anomalies at different granularities are often inherently correlated, where abnormal nodes, edges, and graphs can provide complementary evidence for one another. Modeling each granularity independently may overlook such cross-level dependencies, leading to suboptimal detection performance.

Unified GAD. To better exploit such cross-level dependencies, unified GAD is introduced as a nascent research direction that jointly models node-, edge-, and graph-level anomaly detection within a single framework. While earlier efforts have involved multiple granularities, such as jointly detecting node and edge anomalies in BOURNE Liu et al. (2024) or generating low-level interpretations for graph anomalies in SIGNET Liu et al. (2023a), unifying anomaly detection across levels is not their central objective. As the first attempt, UniGAD Lin et al. (2024) unifies multi-level detection formats through subgraph sampling and facilitates cross-level knowledge sharing with a GraphStitch network. For model training, UniGAD relies on anomaly supervision at different granularities. To reduce such supervision requirements, UniFORM Song et al. (2025) further develops a self-supervised unified framework that homogenizes multi-level objects into subgraph samples and performs anomaly detection with an energy-based model. Despite these advances, existing approaches may overlook granularity-specific information when unifying task formats and still rely on level-specific mechanisms for anomaly detection, limiting knowledge integration across levels. In contrast, our proposed UniTRACER jointly addresses these limitations by preserving granularity-specific information within a target-context joint modeling framework, while consolidating multi-level anomaly detection under a shared reconstruction-based anomaly detector.

Problem Formulation. Let $\mathcal{D} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_M\}$ denote a collection of attributed graphs, where each graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{X}_i)$ consists of a node set $\mathcal{V}_i$, an edge set $\mathcal{E}_i$, and a node feature matrix $\mathbf{X}_i \in \mathbb{R}^{N_i \times D}$, with $N_i = |\mathcal{V}_i|$. Accordingly, graph anomalies can be detected at three granularities: nodes $v \in \mathcal{V}_i$, edges $e \in \mathcal{E}_i$, and graphs $\mathcal{G}_i \in \mathcal{D}$.

In this paper, we focus on the **unsupervised unified GAD** problem, which can be formulated as follows: Given a collection of unlabeled graphs $\mathcal{D}$, the goal is to learn a unified anomaly detector $f(\cdot)$ that jointly assigns anomaly scores to detection objects at node, edge, and graph levels,

$$s_o = f(o), \qquad o \in \mathcal{O}^n \cup \mathcal{O}^e \cup \mathcal{O}^g, \tag{1}$$

where $\mathcal{O}^n = \bigcup_i \mathcal{V}_i$, $\mathcal{O}^e = \bigcup_i \mathcal{E}_i$, and $\mathcal{O}^g = \mathcal{D}$ denote the node-, edge-, and graph-level detection object sets, respectively. A larger s_o indicates a higher likelihood that object o is anomalous. No anomaly annotations at any granularity are available during training.

3 METHOD

In this section, we propose UniTRACER for unified GAD. As shown in Figure 2, we formulate node-, edge-, and graph-level anomaly detection under a shared target-context joint modeling paradigm: the input format of the model is constructed as a *target-context pair* (Sec. 3.1), and the representation learning objective is formulated as *estimating contextual importance conditioned on the detection target* (Sec. 3.2). The learned target-context collaboration is then characterized as shared detection evidence, based on which multi-level anomaly detection is formulated as a unified *evidence reconstruction* task (Sec. 3.3).

3.1 UNIFIED INPUT FORMAT: TARGET-CONTEXT PAIR

As the first step toward unified GAD, heterogeneous detection objects need to be transformed into a common input format. Existing unified GAD methods typically convert different detection objects into subgraph representations for unified modeling. Although this homogenization provides a shared

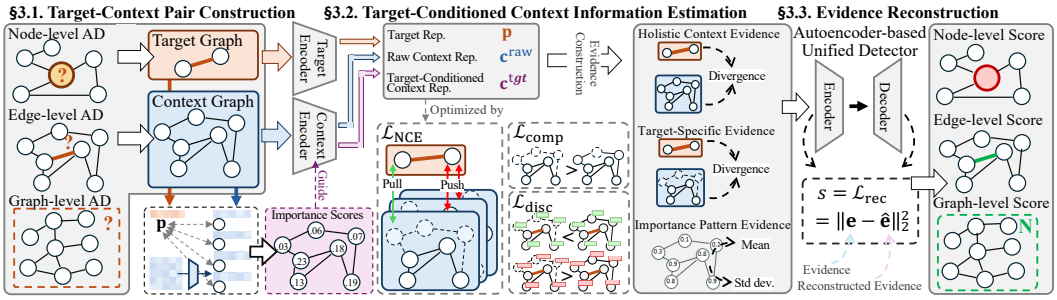


Figure 2: The overall framework of UniTRACER.

input interface across different granularities, it may obscure the granularity-specific information required by individual detection tasks. To address this limitation, we establish a unified input in a more expressive format: a *target-context pair*, where the target graph specifies the object to be detected, along with a context graph that provides its surrounding graph information.

Specifically, the target graph $\mathcal{G}^{\mathrm{t}} = (\mathcal{V}^{\mathrm{t}}, \mathcal{E}^{\mathrm{t}}, X^{\mathrm{t}})$ specifies the object to be assessed, which directly aligns with the current detection objective. In this way, the model can explicitly identify the prediction target and use it to guide subsequent representation learning, hence preserving granularity-specific information required by different detection tasks. On the other hand, the abnormality of a detection target usually depends on how it relates to its surrounding graph environment. To capture such contextual information, we construct a context graph $\mathcal{G}^{\mathrm{c}} = (\mathcal{V}^{\mathrm{c}}, \mathcal{E}^{\mathrm{c}}, X^{\mathrm{c}})$, which provides the structural and attribute information associated with the target.

The target-context formulation is instantiated differently according to the detection granularity. For **node-level detection**, the queried node is represented as a singleton graph, while its K-hop neighborhoods form the context graph to provide local supporting information. Similarly, for **edge-level detection**, the two endpoints of the queried edge and their neighbors are utilized as the target and context graph, respectively. When it comes to **graph-level detection**, the entire queried graph serves as the target graph and also provides its own contextual information for global assessment. Therefore, different detection granularities differ only in how the target-context pair is instantiated, while the subsequent pipeline remains fully shared. In summary, the target-context formulation provides a common interface for node-, edge-, and graph-level anomaly detection, while explicitly modeling the granularity-specific objective into the target graph.

3.2 Unified Learning: Target-Conditioned Context Importance Estimation

Given the unified target-context pair as the input format, the next step is to learn discriminative knowledge for anomaly detection from the input graph pairs. In GAD, a common understanding is that the relationship between a target and its context is highly informative for anomaly detection Liu et al. (2021b); Qiao & Pang (2023). Motivated by this, we may directly compare the target and context graphs to characterize anomaly patterns for unified GAD. Nevertheless, not all contextual information contributes equally to a specific detection problem. In practice, different levels of anomaly detection may favor specific contextual patterns that are particularly relevant to the current detection target He et al. (2024); Liu et al. (2023a). This motivates us to characterize how contextual information contributes to different detection targets in unified GAD.

To explicitly emphasize the contextual information relevant to the current target granularity within a unified task format, we estimate context importance based on the current detection target. The estimated importance is then incorporated into context propagation to derive a target-conditioned context representation. In this way, the target-specific information can be extracted to complement the holistic contextual information, thereby fulfilling the granularity-specific preference. As annotations are unavailable, we further incorporate dedicated unsupervised training objectives to encourage the highlighted context to be compact, target-specific, and discriminative for anomaly detection.

Context Importance Estimation. In the first step, we obtain a target representation $\mathbf{p} = \mathrm{MLP}\left(\frac{1}{|\mathcal{V}^{\mathrm{t}}|}\sum_{v_i \in \mathcal{V}^{\mathrm{t}}} \mathbf{x}_i^{\mathrm{t}}\right)$ to summarize the current detection target, serving as a query for identifying

useful contextual information. We then use $\mathbf{p}$ to estimate an importance score $I_i = \frac{\exp\left(\mathbf{p}^\top \mathbf{x}_i^{\mathrm{c}}\right)}{\sum_{v_j \in \mathcal{V}^{\mathrm{c}}} \exp\left(\mathbf{p}^\top \mathbf{x}_j^{\mathrm{c}}\right)}$ for each context node $v_i \in \mathcal{V}^{\mathrm{c}}$. As I_i measures how v_i is related to the current detection target, the same contextual information can contribute differently according to different target graphs. This allows the model to capture granularity-specific contextual requirements within a unified task format.

Based on the estimated importance, we inject it into message passing, such that information that is more relevant to the current target contributes more strongly to the target-conditioned context representation. Specifically, we employ an importance-aware GNN over the context graph:

$$\mathbf{h}_i^{\mathrm{c},(l+1)} = \sigma\left(\mathbf{W}_{\mathrm{self}}^{(l)} \mathbf{h}_i^{\mathrm{c},(l)} + \sum_{v_j \in \mathcal{N}(v_i)} \alpha_{ij} \mathbf{W}_{\mathrm{ctx}}^{(l)} \mathbf{h}_j^{\mathrm{c},(l)}\right), \tag{2}$$

where $\mathbf{h}_i^{\mathrm{c},(0)} = \mathbf{x}_i^{\mathrm{c}}$, $\alpha_{ij} = \frac{\exp(I_j)}{\sum_{k \in \mathcal{N}(v_i)} \exp(I_k)}$ is the neighbor-normalized importance, and $\mathbf{W}_{\mathrm{self}}^{(l)}$ and $\mathbf{W}_{\mathrm{ctx}}^{(l)}$ are learnable transformation matrices. Through importance-aware propagation, contextual information is selectively aggregated according to the current detection target, thereby preserving the most relevant contextual cues. After L propagation layers, a readout function is applied to produce the target-conditioned context representation $\mathbf{c}^{\mathrm{tgt}}$. Compared to the raw context representation $\mathbf{c}^{\mathrm{raw}}$, which aggregates contextual information uniformly to encode holistic information, $\mathbf{c}^{\mathrm{tgt}}$ explicitly reflects the contextual information emphasized for the current detection target, thereby capturing granularity-specific information within a unified task format.

Once the representations are established, the remaining challenge is how to learn meaningful target-conditioned importance without anomaly labels. Since direct supervision on context relevance is unavailable, we instead consider what properties the learned importance should exhibit for effective context modeling. Intuitively, it should preserve contextual information relevant to the current task, concentrate on a compact set of informative cues, and remain discriminative for GAD purposes. Accordingly, we introduce three complementary losses to encourage these properties.

Target-Context Alignment Loss. To ensure the favored context remains relevant to the current detection task, we maximize the agreement between the task representation and its target-conditioned context representation. Specifically, we adopt an InfoNCE objective that contrasts the matched target-context pair against contexts from other instances:

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{B} \sum_{k=1}^{B} \log \frac{\exp\left(\mathrm{sim}(\mathbf{p}_k, \mathbf{c}_k^{\mathrm{tgt}})/\tau\right)}{\sum_{k'=1}^{B} \exp\left(\mathrm{sim}(\mathbf{p}_k, \mathbf{c}_{k'}^{\mathrm{tgt}})/\tau\right)}, \tag{3}$$

where $\mathrm{sim}(\cdot,\cdot)$ is implemented with cosine similarity and B is the batch size. Optimizing $\mathcal{L}_{\mathrm{NCE}}$ encourages the target-conditioned context representation to preserve information that is strongly associated with the current target graph, thereby preventing target-irrelevant context from dominating the resulting representation.

Context Compactness Loss. To encourage the importance estimator to concentrate on a limited set of contextual cues, we introduce a compactness regularization on the learned importance distribution. Since the normalized importance weights α_{ji} form a probability distribution, the dispersion of the learned importance can be naturally quantified by its entropy. We therefore promote a compact importance distribution by minimizing its entropy:

$$\mathcal{L}_{\mathrm{comp}} = -\frac{1}{B} \sum_{k=1}^{B} \frac{1}{|\mathcal{V}_k^{\mathrm{c}}|} \sum_{i \in \mathcal{V}_k^{\mathrm{c}}} \sum_{j \in \mathcal{N}_{\mathcal{G}_k^{\mathrm{c}}}(i)} \alpha_{ji} \log \alpha_{ji}. \tag{4}$$

By minimizing $\mathcal{L}_{\mathrm{comp}}$, we encourage the importance distribution to concentrate on a limited set of contextual cues, thereby preventing the highlighted contextual information from being dispersed.

Context Discriminability Loss. To encourage the target-conditioned context representation to remain discriminative for GAD purposes, the context importance should not only preserve target-related information, but also emphasize cues that help to distinguish anomalies from normals. However, without anomaly labels, such discriminability cannot be directly learned. We therefore use

target-context consistency as a surrogate: For a benign target-context pair, the estimated importance should reinforce the contextual cues that are consistent with the target. In contrast, when the context is abnormal with respect to the target, the estimator should not produce the same reinforcement. We construct such pseudo-anomalies by randomly shuffling the context graphs. Accordingly, we define:

$$\mathcal{L}_{\text{disc}} = \frac{1}{B}\sum_{k=1}^{B} \text{ReLU}_m\left(\text{sim}(\mathbf{p}_k, \mathbf{c}_k^{\text{raw}}) - \text{sim}(\mathbf{p}_k, \mathbf{c}_k^{\text{tgt}})\right) + \text{ReLU}_m\left(\text{sim}(\mathbf{p}_k, \bar{\mathbf{c}}_k^{\text{tgt}}) - \text{sim}(\mathbf{p}_k, \bar{\mathbf{c}}_k^{\text{raw}})\right), \tag{5}$$

where $\bar{\mathbf{c}}_k^{\text{raw}}$ and $\bar{\mathbf{c}}_k^{\text{tgt}}$ denote the raw and conditioned representations obtained from the shuffled context graph, respectively. The margin-based ReLU function $\text{ReLU}_m(x) = \max(0, x + m)$ helps to stabilize optimization against noise, making it robust to unsupervised training.

Together, the target-context formulation and target-conditioned importance estimation allow UniTRACER to preserve granularity-specific information within a unified task format. Rather than assigning a fixed contextual representation to different anomaly levels, the model adapts the supporting information according to each detection target. The resulting target-context relation and importance patterns further provide a common form of detection evidence, which we use next to unify the detection objective across anomaly levels.

3.3 Unified Detection Task: Evidence Reconstruction

While target-context joint modeling captures the information required by different detection granularities, the next question is how to formulate anomaly detection under a unified objective. Existing methods typically rely on task-specific detection mechanisms for different granularities, which may fragment anomaly detection knowledge across levels and hinder end-to-end knowledge sharing. In order to overcome this fragmentation, a more promising solution is to formulate multi-level anomaly detection as a single task with a unified detection objective. To this end, we first unify the target-context representations and importance into a shared evidence space, and then formulate multi-level anomaly detection as an evidence reconstruction task under a unified objective.

Shared Detection Evidence. Although target-context joint modeling provides granularity-specific representations for different detection tasks, these representations may exhibit different meanings and distributions across node, edge, and graph levels, making them difficult to be directly modeled by a single detector. To facilitate a unified detection task, we transform the learned target-context information into a shared evidence space with consistent interpretation across detection levels.

To be concrete, the constructed evidence space consists of three complementary components. ❶ **Holistic relation evidence** characterizes the relevance between the target graph and its raw context, thereby providing holistic information without target-conditioned reweighting. ❷ **Target-specific evidence** measures their relation after target-relevant contextual information has been emphasized, highlighting any granularity-specific information. ❸ **Importance pattern evidence** summarizes how contextual importance is distributed for the current target, describing the underlying context-selection pattern beyond the compressed representation. They can be defined as:

$$\text{e}_{\text{raw}} = -\,\text{sim}(\mathbf{p}, \mathbf{c}^{\text{raw}}), \qquad \text{e}_{\text{tgt}} = -\,\text{sim}\big(\mathbf{p}, \mathbf{c}^{\text{tgt}}\big), \qquad \text{e}_{\text{imp}} = -\,[\text{mean}(\mathbf{I}),\, \text{std}(\mathbf{I})], \tag{6}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and $\mathbf{I} = \{I_i\}_{v_i \in \mathcal{V}^c}$ denotes learned contextual importance scores. Although the original detection objects differ across anomaly levels, these three pieces of evidence can be consistently extracted from every target-context pair. Finally, we concatenate them into a shared evidence vector, $\mathbf{e} = [\text{e}_{\text{raw}}, \text{e}_{\text{tgt}}, \text{e}_{\text{imp}}]$, which encodes node-, edge-, and graph-level detection knowledge into a common evidence space.

Unified Anomaly Detection as Evidence Reconstruction. With the shared evidence representation, we formulate multi-level anomaly detection under a single reconstruction objective. Specifically, a shared autoencoder jointly learns the normal evidence patterns across node, edge, and graph levels, achieving end-to-end knowledge integration. Then, a reconstruction loss $\mathcal{L}_{\text{rec}}$ is employed for optimization. The overall process is denoted as:

$$\hat{\mathbf{e}} = \text{Dec}\left(\text{Enc}(\mathbf{e})\right), \qquad \mathcal{L}_{\text{rec}} = \|\mathbf{e} - \hat{\mathbf{e}}\|_2^2. \tag{7}$$

At inference time, we directly use the evidence reconstruction error $\|\mathbf{e} - \hat{\mathbf{e}}\|_2^2$ as the unified anomaly score s at inference time, where a larger error indicates a stronger deviation from the evidence patterns learned from the data. Because the same evidence representation and reconstruction objective

Table 1: Anomaly detection performance in terms of AUROC (in percent, mean±std). Highlighted are the results ranked **first**, second, and third. OOM indicates out-of-memory.

	Datasets	Cora	ACM	Flickr	Tolokers	BM-MN	BM-MS
Node-Level	CoLA	90.77±2.70	71.50±3.92	55.55±1.44	52.30±0.42	80.65±1.09	83.39±1.83
	ANEMONE	92.60±1.28	78.19±1.51	69.38±1.10	54.07±0.25	73.55±3.12	72.56±3.27
	PREM	93.95±0.52	75.03±0.83	70.32±1.35	47.38±8.91	74.27±1.11	86.10±0.09
	TAM	89.15±0.92	OOM	65.46±0.58	OOM	78.99±0.68	84.64±0.66
	FreeGAD	85.03±0.00	77.59±0.00	65.18±0.00	62.18±0.00	72.25±0.00	81.53±0.00
	DIFFGAD	89.69±0.03	OOM	OOM	49.52±0.66	77.22±2.95	79.66±2.33
	BOURNE	91.20±1.69	60.74±0.52	58.00±0.00	56.45±0.78	62.91±0.23	67.97±0.54
	SIGNET	OOM	OOM	OOM	OOM	74.67±0.96	67.45±1.74
	UniForm	91.81±1.67	63.23±0.70	68.08±8.36	58.02±0.69	70.57±1.82	62.94±2.89
	UniTRACER	94.13±0.43	77.56±0.20	72.91±0.61	63.85±1.32	81.67±7.21	87.27±6.61
Edge-Level	SEAL	91.40±0.39	62.98±0.52	OOM	55.03±2.10	63.03±1.65	69.40±0.38
	RGSE	93.08±2.50	OOM	63.03±0.86	64.99±0.62	66.03±0.93	79.92±0.96
	BOURNE	89.31±0.55	69.72±0.56	60.00±0.00	63.68±0.69	69.54±0.27	61.86±0.29
	UniForm	92.79±3.31	67.67±0.76	57.15±0.17	64.57±0.34	74.49±1.31	68.21±1.01
	UniTRACER	94.72±0.54	77.83±0.22	64.59±1.92	67.97±1.12	83.22±5.88	87.60±6.70
	Datasets	BZR	DHFR	PROTEINS	IMDB-B	BM-MN	BM-MS
Graph-Level	CVTGAD	72.09±4.61	63.38±4.43	76.66±1.36	69.73±1.51	71.20±4.44	61.64±4.17
	OCGTL	77.79±0.45	63.42±1.18	72.56±2.33	65.93±0.60	76.66±1.53	71.28±5.21
	FANFOLD	75.99±1.22	63.00±4.15	71.74±1.49	68.17±1.63	63.28±4.88	65.23±5.42
	GLADPro	71.86±0.62	62.93±1.15	77.72±1.81	68.46±1.74	63.84±1.97	64.70±1.07
	HC-GLAD	75.53±1.17	61.36±1.01	72.22±1.49	75.02±1.98	77.49±5.45	62.32±2.31
	MUSE	69.76±1.71	60.17±1.12	78.33±0.28	69.31±0.37	82.17±3.61	67.90±1.98
	SIGNET	76.29±0.90	72.55±1.71	75.56±2.84	73.79±2.81	71.58±1.04	65.07±1.89
	UniForm	67.98±3.84	74.07±0.60	68.25±4.24	63.39±0.37	68.49±2.30	70.11±3.34
	UniTRACER	79.33±1.01	83.43±3.47	91.26±1.48	86.05±2.10	79.47±2.70	77.28±1.12

are shared across all anomaly levels, node-, edge-, and graph-level detection no longer require separate prediction heads or level-specific detection objectives. By unifying multi-level anomaly detection under a unified detection objective, UniTRACER can effectively integrate anomaly detection knowledge across granularities, thereby providing a stronger solution to unsupervised unified GAD.

We train UniTRACER end-to-end by jointly optimizing the evidence reconstruction objective and three objectives for target-conditioned context importance:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{NCE}} + \beta\mathcal{L}_{\text{comp}} + \gamma\mathcal{L}_{\text{disc}}, \tag{8}$$

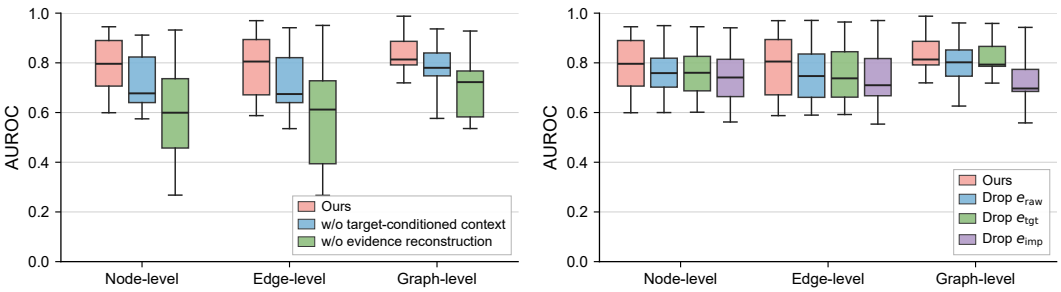
where α, β, and γ are the trade-off parameters. The detailed algorithmic description and detailed complexity analysis are provided in Appendix B and C, respectively.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate UniTRACER and baselines and compare their performance on 17 representative datasets for node-, edge-, and graph- level anomaly detection, including citation networks (Cora, Citeseer, and ACM), social networks (Flickr, Reddit, Tolokers, and IMDB-BINARY), molecular graphs (AIDS, BZR, DHFR, MUTAG, PTC_MR, and NCI1), biochemical graphs (DD and PROTEINS), and pattern recognition graphs (BM-MN and BM-MS) Ding et al. (2019); Tang et al. (2023); Morris et al. (2020a); Liu et al. (2023a); Platonov et al. (2023). Collectively, they span a diverse range of application domains and contain both real-world anomalies and artificially injected anomalies constructed following the protocols in Liu et al. (2021b; 2023a), thereby enabling a rigorous and comprehensive evaluation. For more detailed information, please refer to Appendix D.1.

Baselines and Evaluation Metrics. We compare UniTRACER with 17 representative baselines for node-, edge-, and graph-level GAD. For node-level GAD, we select CoLA Liu et al. (2021b), ANEMONE Jin et al. (2021), PREM Pan et al. (2023), TAM Qiao & Pang (2023), FreeGAD Zhao et al. (2025), DIFFGAD Li et al. (2025), BOURNE Liu et al. (2024), and SIGNET Liu et al. (2023a)



(a) Ablation on model components.

(b) Ablation on descriptors.

Figure 3: Results of ablation study in AUROC (%).

for comparison. For edge-level GAD, we list SEAL Zhang & Chen (2018), RGSE Liu et al. (2023c), and BOURNE Liu et al. (2024) as baselines. For graph-level GAD, our selected baselines include CVTGAD Li et al. (2023), OCGTL Qiu et al. (2022), FanFOLD Cao et al. (2024), GLADPro Yang et al. (2025), HC-GLAD Fu et al. (2026), MUSE Kim et al. (2024), and SIGNET Liu et al. (2023a). Notably, BOURNE naturally supports both node- and edge-level GAD, while SIGNET is extended from graph-level to node-level detection through subgraph-based task adaptation. Moreover, we compare UniTRACER with UniFORM Song et al. (2025), the state-of-the-art unsupervised unified GAD method. Following Song et al. (2025), we adopt AUROC and AUPRC as evaluation metrics. We report average performance and standard deviation over five random runs. More implementation details and hyperparameters used in our experiments are presented in Appendix D.2.

4.2 Experimental Results

Performance Comparison. Table 1 compares UniTRACER with baseline methods in terms of AUROC. Results on more datasets, baselines, and the additional metric AUPRC are reported in Appendix E.1. We make the following observations. ❶ UniTRACER demonstrates strong anomaly detection capabilities in the unsupervised unified GAD setting. As shown in Table 1, UniTRACER achieves state-of-the-art AUROC performance in 16 out of 18 dataset-granularity combinations, while still delivering competitive performance on ACM for node-level GAD and BM-MN for graph-level GAD. ❷ UniTRACER obtains consistent advantages across anomaly granularities. Notably, compared with the unsupervised unified GAD baseline UniForm, UniTRACER improves AUROC by around 10% across all three levels. With the help of task-context joint modeling and evidence reconstruction, UniTRACER can better integrate multi-level anomaly detection knowledge into a unified detection objective. ❸ UniTRACER also shows good memory efficiency. While many baselines encounter OOM issues in large graphs, UniTRACER remains computationally tractable, providing better practical scalability for unified GAD.

Ablation Study on Model Components. To estimate the contribution of each module, we examine the performance by removing either the target-conditioned context information estimation module or the evidence reconstruction module. As shown in Figure 3a, both modules contribute substantially to the overall performance, demonstrating the effectiveness of our design. Notably, when the evidence reconstruction module is removed, anomaly score flipping can happen in node- and edge-level anomaly detection tasks. Nevertheless, the evidence reconstruction module can effectively eliminate this issue to maintain robust GAD results. More ablation experiments on the unified detection objective are in Appendix E.2.

Ablation Study on Evidence Dimensions. To quantify the contribution of each evidence dimension, we compare the full model with variants by dropping one (Figure 3b) or only using one evidence (Appendix E.3). The results show that using all evidence components achieves the best overall performance, and using more evidence generally leads to better results. These results demonstrate that the three evidence dimensions complement each other in capturing different aspects of anomaly semantics. Among them, the importance pattern evidence ($\mathrm{e}_{\mathrm{imp}}$) contributes particularly to graph-level anomaly detection, where removing it results in a substantial performance drop. This is because graph-level anomalies are often determined by a small number of discriminative local structures Liu et al. (2023a), for which explicitly modeling context importance is particularly beneficial.

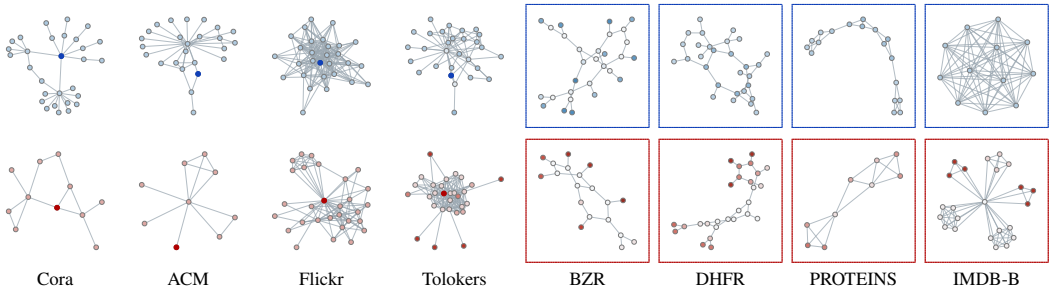


Figure 5: Visualization of context importance. Normal and anomalous examples are shown in blue and red, respectively. Target graphs are indicated by shaded nodes, or boxes for graph-level tasks.

Efficiency Analysis. To examine the efficiency of UniTRACER, we compare the training and inference time on the Cora dataset. As shown in Figure 4 and Appendix E.4, UniTRACER achieves competitive computational efficiency for both training and inference. In particular, it is substantially faster than computationally intensive baselines, while introducing only moderate overhead compared with lightweight methods in exchange for stronger detection performance. Notably, despite its considerable performance improvements, the runtime of UniTRACER remains comparable to UniForm. Overall, these results demonstrate that UniTRACER achieves a favorable trade-off between performance and efficiency.

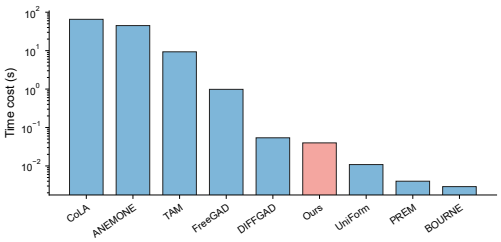


Figure 4: Inference time on Cora.

Visualization of Context Importance. We visualize the learned context importance to examine how UniTRACER adapts to different target graphs. As shown in Figure 5, different target granularities favor distinct target-specific context patterns. For node-level tasks, UniTRACER generally attends to broader neighborhoods, while on Tolokers, anomalous targets emphasize low-degree leaf nodes. For graph-level tasks, the favored context is domain-dependent. For example, on small molecular datasets such as BZR and DHFR, peripheral functional structures are informative for distinguishing both normal and anomalous graphs. To analyze the performance across granularity levels, we further conduct a case study on BM_MN. As shown in Figure 6, the learned context importance reveals a hierarchical correspondence pattern. In particular, the favored context highlighted at the node and edge levels collectively form the broader structural patterns captured at the graph level, indicating effective knowledge integration across granularities. Overall, these results demonstrate the effectiveness of target-conditioned context importance estimation. Additional visualization on representation divergence can be found in Appendix E.5

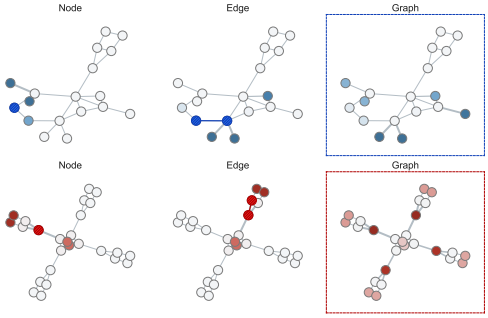


Figure 6: Visualization of context importance.

5 Conclusion

In this paper, we propose a novel unified GAD method termed UniTRACER, which preserves granularity-specific information under a general task formulation and unifies multi-level anomaly detection with a shared objective. To retain task-relevant anomaly cues, we introduce target-context joint modeling with target-conditioned context importance. We further transform target-context information into shared detection evidence and reconstruct it with a shared autoencoder for cross-level knowledge integration. Comprehensive experiments demonstrate the superior performance of UniTRACER over baselines across diverse datasets. The main **limitation** of UniTRACER is its focus on the unsupervised setting, without considering supervision when annotations are available. Therefore, a promising **future direction** is to extend it to supervised or semi-supervised unified GAD.

AI USE STATEMENT

Generative AI tools were used to improve the readability and clarity of the manuscript and to provide suggestions on presentation and organization. We also used Cursor to refactor the code, improving its readability while ensuring that the content was properly anonymized. They were not used to formulate the research hypotheses, design the methodology or experiments, or interpret the experimental results. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHIC STATEMENT

This research does not involve human participants, animal experiments, or the use of sensitive or personally identifiable data. All experiments are conducted using publicly available datasets in controlled experimental settings. We do not identify any specific ethical risks or conflicts of interest associated with this study. Nevertheless, potential real-world deployment of the proposed unified GAD method should account for broader ethical considerations, particularly data privacy, fairness, and the consequences of false positive detections. Appropriate safeguards should therefore be implemented to minimize the risk of unintended bias, discrimination, or other adverse impacts.

REPRODUCIBILITY STATEMENT

All real-world datasets used in our experiments are publicly available, with their corresponding sources and references provided in the paper. To facilitate reproducibility, we provide the detailed introduction and references to datasets in Appendix D.1, hyperparameter search space in Appendix D.2.1, evaluation metrics in Appendix D.2.2 and hardware configuration in Appendix D.2.3. A zipped anonymous code is uploaded in the supplementary material and will be made publicly available upon publication to enable independent reproduction and further research.

REFERENCES

Rui Cao, Shijie Xue, Jindong Li, Qi Wang, and Yi Chang. Fanfold: Graph normalizing flows-driven asymmetric network for unsupervised graph-level anomaly detection. *arXiv preprint arXiv:2407.00383*, 2024.

Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM international conference on data mining*, pp. 594–602. SIAM, 2019.

Paul D Dobson and Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology*, 330(4):771–783, 2003.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 315–324, 2020.

Dongsheng Duan, Lingling Tong, Yangxi Li, Jie Lu, Lei Shi, and Cheng Zhang. Aane: Anomaly aware network embedding for anomalous link detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1002–1007. IEEE, 2020.

Yali Fu, Jindong Li, Jiahong Liu, Qianli Xing, Qi Wang, and Irwin King. Hc-glad: Dual hyperbolic contrastive learning for unsupervised graph-level anomaly detection. *Neural Netw.*, 202(C), August 2026. ISSN 0893-6080. doi: 10.1016/j.neunet.2026.109009. URL https://doi.org/10.1016/j.neunet.2026.109009.

Yuan Gao, Xiang Wang, Xiangnan He, Zhenguang Liu, Huamin Feng, and Yongdong Zhang. Addressing heterophily in graph anomaly detection: A perspective of graph spectrum. In *Proceedings of the ACM web conference 2023*, pp. 1528–1538, 2023.

Dongmei Han, Min Min, Yuchen Wang, Guoming Xu, and Xiaofeng Zhou. Beyond single transactions: D-emaml—dual-edge motif neural networks for enhanced anti-money laundering detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14792–14801, 2026.

Junwei He, Qianqian Xu, Yangbangyan Jiang, Zitai Wang, and Qingming Huang. Ada-gad: Anomaly-denoised autoencoders for graph anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8481–8489, 2024.

Mingguo He, Zhewei Wei, Hongteng Xu, et al. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. *Advances in neural information processing systems*, 34:14239–14251, 2021.

Christoph Helma, Ross D. King, Stefan Kramer, and Ashwin Srinivasan. The predictive toxicology challenge 2000–2001. *Bioinformatics*, 17(1):107–108, 2001.

Ming Jin, Yixin Liu, Yu Zheng, Lianhua Chi, Yuan-Fang Li, and Shirui Pan. Anemone: Graph anomaly detection with multi-scale contrastive learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 3122–3126, 2021.

Sunwoo Kim, Soo Yong Lee, Fanchen Bu, Shinhwan Kang, Kyungho Kim, Jaemin Yoo, and Kijung Shin. Rethinking reconstruction-based graph-level anomaly detection: Limitations and a simple remedy. *Advances in Neural Information Processing Systems*, 37:95931–95962, 2024.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1269–1278, 2019.

Ao Li, Zhou Qin, Runshi Liu, Yiqun Yang, and Dong Li. Spam review detection with graph convolutional networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 2703–2711, 2019.

Jindong Li, Qianli Xing, Qi Wang, and Yi Chang. Cvtgad: Simplified transformer with cross-view attention for unsupervised graph-level anomaly detection. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 185–200. Springer, 2023.

Jinghan Li, Yuan Gao, Jinda Lu, Junfeng Fang, Congcong Wen, Hui Lin, and Xiang Wang. Diffgad: A diffusion-based unsupervised graph anomaly detector. In *International Conference on Learning Representations*, volume 2025, pp. 37744–37766, 2025.

Yiqing Lin, Jianheng Tang, Chenyi Zi, H. Vicky Zhao, Yuan Yao, and Jia Li. UniGAD: Unifying multi-level graph anomaly detection. In *NeurIPS*, 2024.

Jie Liu, Mengting He, Xuequn Shang, Jieming Shi, Bin Cui, and Hongzhi Yin. Bourne: Bootstrapped self-supervised learning framework for unified graph anomaly detection. In *ICDE*, pp. 2820–2833. IEEE, 2024.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: a gnn-based imbalanced learning approach for fraud detection. In *Proceedings of the web conference 2021*, pp. 3168–3177, 2021a.

Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6):2378–2392, 2021b.

Yixin Liu, Kaize Ding, Qinghua Lu, Fuyi Li, Leo Yu Zhang, and Shirui Pan. Towards self-interpretable graph-level anomaly detection. In *NeurIPS*, volume 36, pp. 8975–8987, 2023a.

Yixin Liu, Shirui Pan, Yu Guang Wang, Fei Xiong, Liang Wang, Qingfeng Chen, and Vincent CS Lee. Anomaly detection in dynamic graphs via transformer. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12081–12094, 2023b. doi: 10.1109/TKDE.2021.3124061.

Zhen Liu, Wenbo Zuo, Dongning Zhang, and Xiaodong Feng. Rgse: Robust graph structure embedding for anomalous link detection. *IEEE Transactions on Big Data*, 9(5):1420–1429, 2023c.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *CoRR*, abs/2007.08663, 2020a. URL https://arxiv.org/abs/2007.08663.

Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020b.

Junjun Pan, Yixin Liu, Yizhen Zheng, and Shirui Pan. Prem: A simple yet effective approach for node-level graph anomaly detection. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1253–1258, 2023. doi: 10.1109/ICDM58522.2023.00157.

Junjun Pan, Yu Zheng, Yue Tan, and Yixin Liu. A survey of generalization of graph anomaly detection: From transfer learning to foundation models. *arXiv preprint arXiv:2509.06609*, pp. 316–323, November 2025. doi: 10.1109/ICKG66886.2025.00048. URL https://doi.ieeecomputersociety.org/10.1109/ICKG66886.2025.00048.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at evaluation of gnns under heterophily: Are we really making progress? In *The Eleventh International Conference on Learning Representations*, 2023.

Hezhe Qiao and Guansong Pang. Truncated affinity maximization: One-class homophily modeling for graph anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Chen Qiu, Marius Kloft, Stephan Mandt, and Maja Rudolph. Raising the bar in graph-level anomaly detection. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2196–2203. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/305. URL https://doi.org/10.24963/ijcai.2022/305. Main Track.

Chuancheng Song, Xixun Lin, Hanyang Shen, Yanmin Shang, and Yanan Cao. Uniform: Towards unified framework for anomaly detection on graphs. In *AAAI*, volume 39, pp. 12559–12567, 2025.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International conference on machine learning*, pp. 21076–21089. PMLR, 2022.

Jianheng Tang, Fengrui Hua, Ziqi Gao, Peilin Zhao, and Jia Li. Gadbench: Revisiting and benchmarking supervised graph anomaly detection. *Advances in Neural Information Processing Systems*, 36:29628–29653, 2023.

S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.

Nikil Wale, Ian A Watson, and George Karypis. Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems*, 14(3):347–375, 2008.

Yili Wang, Yixin Liu, Xu Shen, Chenyu Li, Rui Miao, Kaize Ding, Ying Wang, Shirui Pan, and Xin Wang. Unifying unsupervised graph-level anomaly detection and out-of-distribution detection: A benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 91903–91929, 2025.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Yiming Xu, Zihan Chen, Zhen Peng, Song Wang, Bin Shi, Bo Dong, and Chao Shen. Generalist graph anomaly detection via prototype-based distillation. *arXiv preprint arXiv:2605.26857*, 2026.

Zhenyu Yang, Ge Zhang, Jia Wu, Jian Yang, Shan Xue, Amin Beheshti, Hao Peng, and Quan Z Sheng. Global interpretable graph-level anomaly detection via prototype. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3586–3597, 2025.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31, 2018.

Yunfeng Zhao, Yixin Liu, Shiyuan Li, Qingfeng Chen, Yu Zheng, and Shirui Pan. Freegad: A training-free yet effective approach for graph anomaly detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4379–4389, 2025.

Li Zheng, Zhenpeng Li, Jian Li, Zhao Li, and Jun Gao. Addgraph: Anomaly detection in dynamic graph using attention-based temporal gcn. In *IJCAI*, volume 3, pp. 7, 2019.

A Detailed Related Work

A.1 Graph Anomaly Detection

Graph anomaly detection (GAD) aims to identify abnormal objects in graph-structured data, where abnormality can manifest at node, edge, and graph levels. As noted in the main paper, most existing approaches are granularity-specific to a particular detection level. In the following, we provide a detailed Description of each level.

A.1.1 Node-level Anomaly Detection

Node-level anomaly detection, as the most extensively investigated setting in GAD, focuses on identifying individual nodes that deviate from the majority. Existing approaches can be broadly categorized into unsupervised and supervised paradigms.

Unsupervised paradigms. Reconstruction-based methods learn node representations by reconstructing the graph structure and node attributes, where nodes with high reconstruction errors are flagged as anomalies. DOMINANT Ding et al. (2019) employs a graph convolutional encoder with separate structure and attribute reconstruction decoders, computing anomaly scores based on reconstruction mean squared error. Beyond reconstruction, contrastive learning has emerged as another prominent paradigm. CoLA Liu et al. (2021b) is the first contrastive self-supervised framework for GAD, sampling node-subgraph instance pairs and deriving anomaly scores from their agreement. PREM Pan et al. (2023) further simplifies this paradigm by introducing a preprocessing module and an ego-neighbor matching mechanism, eliminating the need for message-passing propagation during training. Another line of work focuses on affinity modeling. TAM Qiao & Pang (2023), for instance, learns node representations by maximizing local affinity on truncated graphs where non-homophily edges are iteratively removed. Finally, denoising-based methods, including ADA-GAD He et al. (2024) and DiffGAD Li et al. (2025), incorporate denoising mechanisms or diffusion-based latent space learning to enhance robustness against noisy graph data.

Supervised paradigms. Supervised approaches formulate node-level GAD as a binary classification problem, assuming that labels of both normal and anomalous nodes are available for model training Tang et al. (2023). Therefore, related studies mainly introduce GAD methods under the binary classification paradigm Tang et al. (2022); Dou et al. (2020); Li et al. (2019); Liu et al. (2021a); He et al. (2021); Gao et al. (2023). For example, BWGNN Tang et al. (2022) designs Beta Wavelet Graph Neural Networks with flexible band-pass filters to capture high-frequency anomalies that are typically smoothed out by conventional GNNs.

Despite these advances, node-level methods are inherently limited to detecting point anomalies and cannot capture anomalies that manifest at higher granularities.

A.1.2 EDGE-LEVEL ANOMALY DETECTION

Edge-level anomaly detection focuses on identifying abnormal links that represent unexpected or irregular relationships between entities. This task is more challenging than node-level detection because edge anomalies depend not only on the characteristics of connected nodes but also on the semantic meaning of their interactions. In dynamic graphs, the temporal dynamics of these interactions further complicate the detection. AANE Duan et al. (2020) is a representative framework that explicitly models the existence of anomalous links. It learns node embeddings via GCN and computes link probability using the hyperbolic tangent of these embeddings, with a loss function consisting of an anomaly-aware loss and an adjusted fitting loss. RGSE Liu et al. (2023c) leverages robust graph structure embedding to generate both global and local stable views, facilitating the detection of anomalous links within contaminated graphs. AddGraph Zheng et al. (2019) integrates temporal attention into GCNs to capture evolving connectivity patterns in dynamic graphs, encoding historical activation patterns in a memory module. TADDY Liu et al. (2023b), a Transformer-based framework for dynamic graphs, employs self-attention mechanisms to model coupled spatial-temporal patterns for anomalous link detection. A key limitation of edge-level methods is that they typically treat edge anomalies in isolation, without explicitly leveraging complementary anomaly signals from node- or graph-level detection.

A.1.3 GRAPH-LEVEL ANOMALY DETECTION

Graph-level anomaly detection (GLAD) identifies anomalous graphs from a collection and has recently attracted increasing attention, with various learning paradigms explored. These paradigms input transformation learning, where OCGTL Qiu et al. (2022) combines deep one-class classification with self-supervised graph transformation learning and uses the transformation classification loss as an anomaly score; reconstruction-based methods, where MUSE Kim et al. (2024) argues that taking multifaceted summaries of reconstruction errors beyond the mean can further strengthen features and achieve state-of-the-art performance with a surprisingly simple method; distribution modeling, where FANFOLD Cao et al. (2024) introduces graph normalizing flows to learn the latent distribution of normal graphs, transforming it into a standard normal distribution and detecting anomalies via distributional deviation using an asymmetric network with knowledge distillation; prototype learning, where GLADPro Yang et al. (2025) incorporates prototype learning with the information bottleneck principle to capture the most significant subgraph patterns as global-level explanations, and ProMoS Xu et al. (2026) extends this by distilling knowledge from multiple source domains into prototypes for cross-domain anomaly detection without fine-tuning; and structural representation learning, where CVTGAD Li et al. (2023) employs a simplified Transformer with cross-view attention to explore intra- and inter-graph relationships, while HC-GLAD Fu et al. (2026) constructs hypergraphs based on pre-designed gold motifs and performs hypergraph convolution to capture high-order node group information, introducing hyperbolic geometry with the hyperboloid model to preserve the hierarchical structure of real-world graphs. Despite these advances, graph-level methods operate in isolation and cannot leverage node or edge-level anomaly signals.

A.2 UNIFIED GAD

The observation that anomalies at different granularities are often inherently correlated has motivated the nascent research direction of unified GAD, which aims to jointly model node-, edge-, and graph-level anomaly detection within a single framework. While earlier efforts have involved multiple granularities, such as jointly detecting node and edge anomalies in BOURNE Liu et al. (2024) or generating low-level interpretations for graph anomalies in SIGNET Liu et al. (2023a), unifying anomaly detection across node-, edge-, and graph-levels is not their central objective. UniGAD Lin et al. (2024) is the first unified framework for jointly detecting node-, edge-, and graph-level anomalies; it employs the Maximum Rayleigh Quotient Subgraph Sampler (MRQSampler) to transform objects at each level into graph-level tasks on subgraphs while maximizing the accumulated spectral energy (i.e., the Rayleigh quotient) with theoretical guarantees, and introduces a GraphStitch Network to integrate cross-level information and harmonize conflicting training goals, though it relies on anomaly supervision at different granularities. UniFORM Song et al. (2025) further develops a self-supervised unified framework that comprises three modules: UIO unifies multi-level tasks from a subgraph perspective using an energy-based GNN, UMC enhances meta-learning via contrastive learning and Langevin dynamics to generate phantom samples as substitutes for anomalous data,

Algorithm 1 Training Procedure of UniTRACER

Input: Training task collection $\mathcal{T}$, number of epochs E, batch size B.
Output: Trained model parameters Θ.

Initialize model parameters Θ and optimizer $\mathcal{O}$.
for $e = 1, \ldots, E$ **do**
  for each level **do**
    Construct mini-batches $\{\mathcal{B}_1, \ldots, \mathcal{B}_M\}$ from $\mathcal{D}$.
    for $\mathcal{B}_m \in \{\mathcal{B}_1, \ldots, \mathcal{B}_M\}$ **do**
      Construct target-context graph pairs $\{(\mathcal{G}_i^{\mathrm{t}}, \mathcal{G}_i^{\mathrm{c}})\}_{i \in \mathcal{B}_m}$.
      Estimate importance score $\{I_i\}$ and obtain target and context graph representations $\mathbf{p}$, $\mathbf{c}^{\mathrm{raw}}$, and $\mathbf{c}^{\mathrm{tgt}}$.
      Construct evidence $\mathbf{e}$ and obtain reconstructioned evidence $\hat{\mathbf{e}} = \mathrm{Dec}(\mathrm{Enc}(\mathbf{e}))$.
      Compute $\mathcal{L}_{\mathrm{rec}}$, $\mathcal{L}_{\mathrm{NCE}}$, $\mathcal{L}_{\mathrm{comp}}$, and $\mathcal{L}_{\mathrm{disc}}$ following their corresponding definitions.
      $\mathcal{L}_{\mathcal{B}_m} \leftarrow \mathcal{L}_{\mathrm{rec}} + \alpha\mathcal{L}_{\mathrm{NCE}} + \beta\mathcal{L}_{\mathrm{comp}} + \gamma\mathcal{L}_{\mathrm{disc}}$.
      $\mathcal{L}_{\mathcal{B}_m} \leftarrow \mathcal{L}_{\mathcal{B}_m}/|\mathcal{D}|$. *// Normalize loss scales across task levels.*
      Update Θ with optimizer $\mathcal{O}$ using $\mathcal{L}_{\mathcal{B}_m}$.
    end for
  end for
end for
return trained UniTRACER model.

and UPL designs a unified loss from intra-inter perspectives; through these designs, UniFORM homogenizes multi-level objects into subgraph samples and performs detection with an energy-based model. Despite these advances, existing approaches may overlook granularity-specific information when unifying task formats and still rely on level-specific mechanisms for anomaly detection, limiting knowledge integration across levels. In contrast, our proposed UniTRACER jointly addresses these limitations by preserving granularity-specific information within a target-context joint modeling framework, while consolidating multi-level anomaly detection under a shared reconstruction-based anomaly detector.

B Algorithm

In this section, we summarize the training procedure of UniTRACER in Algorithm 1. In brief, we perform unsupervised training on all datasets, with node-level and graph-level tasks using mini-batches to reduce memory consumption. Node-level tasks are constructed from nodes in the dataset along with their sampled 2-hop contexts. For graph-level tasks, graphs with excessive numbers of nodes are randomly partitioned into smaller subgraphs for training to avoid excessive memory usage. We normalize the losses from node-level and graph-level tasks to ensure they are on the same scale.

C Complexity Analysis

We analyze the computational complexity of UniTRACER. Let V and E denote the numbers of nodes and edges involved in a training graph, respectively. The overall complexity is therefore bounded by graph-level anomaly detection, where the entire graph is used as both the target and context graph. In this case, for each target-context pair, constructing the target representation and estimating context importance require $\mathcal{O}(V)$ time. The subsequent context encoding is dominated by GNN message passing, which requires $\mathcal{O}(E)$ time, while graph readout and evidence construction incur $\mathcal{O}(V)$ complexity. The remaining evidence reconstruction introduces only small overhead $\mathcal{O}(V)$ for computing importance pattern evidence. Therefore, the overall time complexity of UniTRACER is $\mathcal{O}(V + E)$. Therefore, UniTRACER maintains computational complexity comparable to that of training a vanilla GNN.

D Implementation Details

D.1 Description of Datasets

In this section, we present additional details of all datasets used in our experiments. Table 2 summarizes their statistics and task granularities.

Cora Liu et al. (2021b) is a citation network in which nodes represent scientific papers and edges represent citation relationships. Each node is associated with a bag-of-words feature vector extracted from the paper's abstract, whose dimension is determined by the size of the corresponding dictionary.

Citeseer Liu et al. (2021b) is a citation network in which nodes correspond to scientific papers and edges represent citation relationships. Each node is described by a bag-of-words feature vector, and papers are categorized into six research areas.

ACM Ding et al. (2019) is a citation network constructed from the ACM Digital Library, containing papers as nodes and citation links as edges. Node attributes are bag-of-words representations of paper contents.

Flickr Ding et al. (2019) is a social network derived from the photo-sharing platform, where nodes represent users and edges denote following or follower relationships. The dataset contains approximately 7,575 nodes and 239,738 edges, with a small portion of nodes labeled as anomalies. Node features are constructed from user profile information.

Reddit Kumar et al. (2019) is a social network dataset where nodes represent users and edges denote interaction relationships. It contains 10,984 nodes and 168,016 edges. The dataset provides ground-truth node-level anomaly labels, where anomalous users are those banned from the platform. Node features are constructed from user activity and subreddit participation.

Tolokers Platonov et al. (2023) is a social network for crowdsourcing, where nodes represent workers who participated in crowd-sourcing projects, and edges connect two workers who worked on the same task. The graph contains 11,758 nodes and 530,758 edges. The task is to predict which workers have been banned, using profile and task performance features.

IMDB-BINARY is a social network for movie collaboration, consisting of the ego-networks of 1,000 actors/actresses. Nodes represent actors/actresses, and an edge connects two nodes if they appear in the same movie. The graph label is binary, indicating whether the corresponding movie genre is Action or Romance.

AIDS Morris et al. (2020b) is a molecular graph dataset from the TU collection, containing 2,000 graphs with an average of 15.59 nodes and 47.98 edges. Nodes represent atoms and edges represent chemical bonds. The original task is graph-level classification of compounds as active or inactive against HIV. For anomaly detection, graphs of one class are treated as normal and the remaining as anomalies.

BZR Wu et al. (2018) is a molecular graph dataset from the TU collection, containing 405 molecular graphs of chemical compounds. Nodes correspond to atoms and edges to chemical bonds. Graph labels distinguish active and inactive ligands of the benzodiazepine receptor.

DHFR Morris et al. (2020b) is a molecular graph dataset from the TU collection, comprising 756 molecular graphs that describe the inhibitory activity of compounds against dihydrofolate reductase, an enzyme critical in cellular folate metabolism. Nodes represent atoms and edges represent chemical bonds.

MUTAG Debnath et al. (1991) is a molecular graph dataset from the TU collection, consisting of 188 graphs with an average of 17.93 nodes and 57.52 edges. Nodes correspond to atoms and edges to chemical bonds. The graph label indicates whether the molecule is mutagenic. Following the standard setting, non-mutagenic molecules are treated as normal and mutagenic molecules as anomalies, with -NO_2 and -NH_2 groups serving as ground-truth explanations.

PTC_MR Helma et al. (2001) is a molecular graph dataset from the TU collection, containing 344 graphs with an average of 14.29 nodes and 29.38 edges. Nodes represent atoms and edges represent

Table 2: Statistics of the datasets.

Datasets	Task granularity	# Graphs	Avg. Nodes	Avg. Edges
Citation Network				
Cora	Node/Edge	1	2708	5429
Citeseer	Node/Edge	1	3327	4552
ACM	Node/Edge	1	16484	71980
Social Networks				
Flickr	Node/Edge	1	7575	239738
Reddit	Node/Edge	1	10984	168016
Tolokers	Node/Edge	1	11758	530758
IMDB-BINARY	Graph	1000	19.77	212.84
Molecular Graphs				
AIDS	Graph	2000	15.59	47.98
BZR	Graph	405	35.75	76.72
DHFR	Graph	756	42.43	89.09
MUTAG	Graph	188	17.93	57.52
PTC_MR	Graph	344	14.29	29.38
NCI1	Graph	4110	29.87	64.60
Biochemistry Graphs				
DD	Graph	1178	284.32	1715.63
PROTEINS	Graph	1113	39.06	145.63
Pattern Recognition Graphs				
BM-MN	Node/Edge/Graph	700	18.44	57.19
BM-MS	Node/Edge/Graph	700	14.04	43.20

chemical bonds. The task is to predict the carcinogenicity of chemical compounds. For anomaly detection, one class is treated as normal and the other as anomalous.

NCI1 Wale et al. (2008) is a molecular graph dataset from the TU collection, comprising 4,110 graphs with an average of 29.87 nodes and 64.60 edges. Nodes correspond to atoms and edges to chemical bonds. The dataset is designed for graph-level classification of compounds based on their anti-cancer activity. In anomaly detection, one class is considered normal and the other anomalous.

DD Vishwanathan et al. (2010) is a biochemistry dataset from the TU collection, containing 1,178 graphs with an average of 284.32 nodes and 1,715.63 edges. Nodes represent amino acids and edges represent spatial proximity in protein structures. The original task is graph-level classification of proteins. For anomaly detection, graphs of one class are treated as normal and the remaining as anomalies.

PROTEINS Dobson & Doig (2003) is a biochemistry dataset from the TU collection, containing protein graphs where nodes represent amino acids and edges represent structural proximity. Each graph is labeled according to whether it is an enzyme or a non-enzyme.

BM-MN Liu et al. (2023a) is a pattern recognition dataset constructed with motif-based anomalies. Each graph consists of a base structure and a varying number of house motifs. Normal graphs contain one or two house motifs, while anomalous graphs contain three or four house motifs. The ground-truth explanations are defined as the nodes and edges within the motifs. This dataset supports both node-level and graph-level anomaly detection.

BM-MS Liu et al. (2023a) is a pattern recognition dataset with motif-size-based anomalies. Each graph consists of a base structure and a cycle motif. Normal graphs contain a cycle motif with 3–5 nodes, while anomalous graphs contain a cycle motif with 6–9 nodes. The ground-truth explanations are the nodes and edges within the cycle motif. This dataset supports both node-level and graph-level anomaly detection.

Table 3: Node- and edge-level anomaly detection (AUROC / AUPRC, in percent, mean±std). Highlighted are the results ranked **first**, second, and third. OOM indicates out-of-memory.

	Datasets	Cora	Citeseer	ACM	Flickr	Reddit	Tolokers
				AUROC			
Node-level	CoLA	90.77±2.70	88.04±2.17	71.50±3.92	55.55±1.44	54.97±2.90	52.30±0.42
	ANEMONE	92.60±1.28	92.35±4.77	78.19±1.51	69.38±1.10	55.35±0.66	54.07±0.25
	PREM	93.95±0.52	94.16±0.64	75.03±0.83	70.32±1.35	53.46±5.98	47.38±8.91
	TAM	89.15±0.92	91.55±0.98	OOM	65.46±0.58	OOM	OOM
	FreeGAD	85.03±0.00	86.34±0.00	77.59±0.00	65.18±0.00	57.21±0.00	62.18±0.00
	DIFFGAD	89.69±0.03	87.74±0.00	OOM	OOM	56.06±0.02	49.52±0.66
	BOURNE	91.20±1.69	90.12±2.23	60.74±0.52	58.00±0.00	56.80±0.31	56.45±0.78
	SIGNET	OOM	OOM	OOM	OOM	OOM	OOM
	UniForm	91.81±1.67	90.98±1.82	63.23±0.70	68.08±8.36	56.79±1.63	58.02±0.69
	Ours	94.13±0.43	94.52±0.90	77.56±0.20	72.91±0.61	59.93±0.94	63.85±1.32
Edge-level	SEAL	91.40±0.39	91.30±0.92	62.98±0.52	OOM	56.55±0.27	55.03±2.10
	RGSE	93.08±2.50	90.19±2.65	OOM	63.03±0.86	52.41±0.36	64.99±0.62
	BOURNE	89.31±0.55	91.14±2.53	69.72±0.56	60.00±0.00	54.76±0.09	63.68±0.69
	UniForm	92.79±3.31	94.45±3.59	67.67±0.76	57.15±0.17	54.86±0.73	64.57±0.34
	Ours	94.72±0.54	96.97±0.31	77.83±0.22	64.59±1.92	58.76±1.65	67.97±1.12
				AUPRC			
Node-level	CoLA	41.80±2.14	35.23±3.09	24.71±4.02	7.22±0.46	3.65±0.31	22.39±0.16
	ANEMONE	47.39±1.64	38.21±5.14	24.82±3.38	14.40±1.60	3.59±0.08	23.45±0.13
	PREM	58.08±1.87	35.71±1.50	12.69±0.18	14.69±0.12	4.27±0.87	20.74±3.80
	TAM	54.90±0.19	34.58±0.20	OOM	15.09±0.09	OOM	OOM
	FreeGAD	50.11±0.00	33.67±0.00	18.87±0.00	3.72±0.00	4.25±0.00	26.16±0.00
	DIFFGAD	45.89±0.02	37.49±0.00	OOM	OOM	3.73±0.00	25.88±0.41
	BOURNE	57.93±1.26	34.29±0.18	16.61±0.02	15.94±0.00	4.74±0.04	22.96±1.10
	SIGNET	OOM	OOM	OOM	OOM	OOM	OOM
	UniForm	55.08±0.45	33.05±0.09	15.62±0.23	16.45±1.19	4.03±0.20	24.26±0.30
	Ours	64.97±4.88	41.93±0.57	22.57±0.60	19.32±5.68	5.13±0.32	27.55±1.19
Edge-level	SEAL	81.85±0.76	81.06±0.96	18.40±0.09	OOM	5.71±0.01	56.01±1.40
	RGSE	83.10±0.78	72.62±0.74	OOM	17.45±0.44	5.96±0.04	56.28±0.15
	BOURNE	74.21±0.06	75.79±0.46	19.28±0.11	16.94±0.00	5.95±0.01	58.68±0.69
	UniForm	78.50±0.55	79.57±0.58	20.56±0.20	16.21±0.03	5.21±0.06	58.49±0.66
	Ours	82.42±3.51	82.36±0.19	24.66±0.64	19.72±2.00	6.67±0.58	63.92±0.93

D.2 Details of Implementation

D.2.1 Hyper-parameters

We tune the key hyper-parameters of UniTRACER via random search over predefined grids. Specifically, the search space inputs the learning rate $\mathrm{lr} \in \{1\times10^{-3},\ 1\times10^{-4}\}$, the temperature coefficient $\tau \in \{0.1,\ 0.2,\ 0.3\}$, and loss coefficients $\alpha, \beta, \gamma \in \{0.0,\ 0.01,\ 0.1,\ 0.5,\ 1.0\}$. We employ the adam optimizer for optimization.

D.2.2 Evaluation Metrics

Following commonly used evaluation protocols in node-, edge-, and graph-level GAD Wang et al. (2025); Song et al. (2025), we assess anomaly detection performance with AUROC (area under the receiver operating characteristic curve) and AUPRC (area under the precision–recall curve). These two metrics characterize detection quality from complementary perspectives, with larger values indicating stronger performance. To account for randomness in model training, we repeat each experiment five times independently and report the average results together with the corresponding standard deviations.

Table 4: Graph-level anomaly detection (AUROC / AUPRC, in percent, mean±std). Highlighted are the results ranked **first**, second, and third. OOM indicates out-of-memory.

Datasets	AIDS	BZR	DHFR	MUTAG	PTC_MR	DD	NCI1	PROTEINS	IMDB-B
AUROC									
CVTGAD	95.47±1.18	72.09±4.61	63.38±4.43	80.06±19.85	54.92±1.24	OOM	67.94±1.08	76.66±1.36	69.73±1.51
OCGTL	97.96±0.82	77.79±0.45	63.42±1.18	84.09±0.29	55.54±0.39	77.46±0.91	74.10±0.46	72.56±2.33	65.93±0.60
FANFOLD	97.51±0.82	75.99±1.22	63.00±4.15	81.42±1.44	61.37±4.25	72.45±1.33	62.46±1.00	71.74±1.49	68.17±1.63
GLADPro	95.97±1.60	71.86±0.62	62.93±1.15	91.14±1.69	57.04±1.82	75.38±1.72	58.08±1.64	77.72±1.81	68.46±1.74
HC-GLAD	97.33±0.24	75.53±1.17	61.36±1.01	89.89±0.75	56.32±1.54	OOM	76.01±0.83	72.22±1.49	75.02±1.98
MUSE	96.34±1.28	69.76±1.71	60.17±1.12	67.93±3.56	56.57±0.69	80.67±0.25	76.13±1.28	78.33±0.28	69.31±0.37
SIGNET	97.45±0.96	76.29±0.90	72.55±1.71	89.71±13.66	57.18±2.06	74.83±0.69	75.02±0.71	75.56±2.84	73.79±2.81
UniForm	94.34±1.84	67.98±3.84	74.07±0.60	75.79±1.82	54.37±0.78	65.72±3.57	61.57±3.71	68.25±4.24	63.39±0.37
Ours	98.25±0.21	79.33±1.01	83.43±3.47	98.77±0.62	71.94±2.94	81.35±1.28	78.99±1.37	91.26±1.48	86.05±2.10
AUPRC									
CVTGAD	96.43±1.21	35.04±1.55	78.45±2.79	88.40±1.26	53.72±1.06	OOM	70.19±1.67	78.94±1.45	65.09±1.67
OCGTL	97.98±0.16	35.39±0.15	81.68±0.77	89.99±0.19	54.79±0.21	68.70±0.33	71.28±0.30	71.62±1.66	64.05±0.41
FANFOLD	97.10±0.40	30.91±2.20	75.77±3.44	88.56±1.85	57.37±2.38	70.85±1.07	64.44±0.52	73.67±1.53	68.18±1.15
GLADPro	96.83±1.14	35.15±1.09	71.16±1.29	91.28±1.63	59.31±1.32	71.81±1.87	59.26±1.88	77.84±1.24	70.29±1.96
HC-GLAD	97.68±1.60	42.29±2.36	67.17±1.40	85.48±0.74	57.78±1.03	OOM	69.57±1.04	74.83±0.84	72.18±1.42
MUSE	96.33±1.22	37.73±1.75	68.86±0.54	85.23±0.29	53.08±0.44	70.20±0.49	72.70±1.03	78.17±0.55	67.35±0.63
SIGNET	96.89±0.97	42.45±0.43	77.54±1.90	83.70±1.14	57.61±1.95	66.29±0.28	70.56±1.28	81.58±3.17	75.26±2.86
UniForm	93.30±1.04	37.33±1.14	83.80±0.93	82.80±2.25	53.30±0.38	60.96±2.10	63.87±1.83	64.29±3.66	66.10±0.19
Ours	98.76±0.24	45.10±2.35	90.36±1.74	99.36±0.36	68.23±3.95	72.28±2.75	74.17±1.47	88.99±2.57	81.07±4.15

D.2.3 Hardware Configuration

All experiments are performed on a Windows-based workstation with an AMD Ryzen 5800X CPU and 64 GB of system memory. Model training and inference are carried out using a single NVIDIA GeForce RTX 4090 GPU equipped with 24 GB of video memory.

E Additional Experimental Results

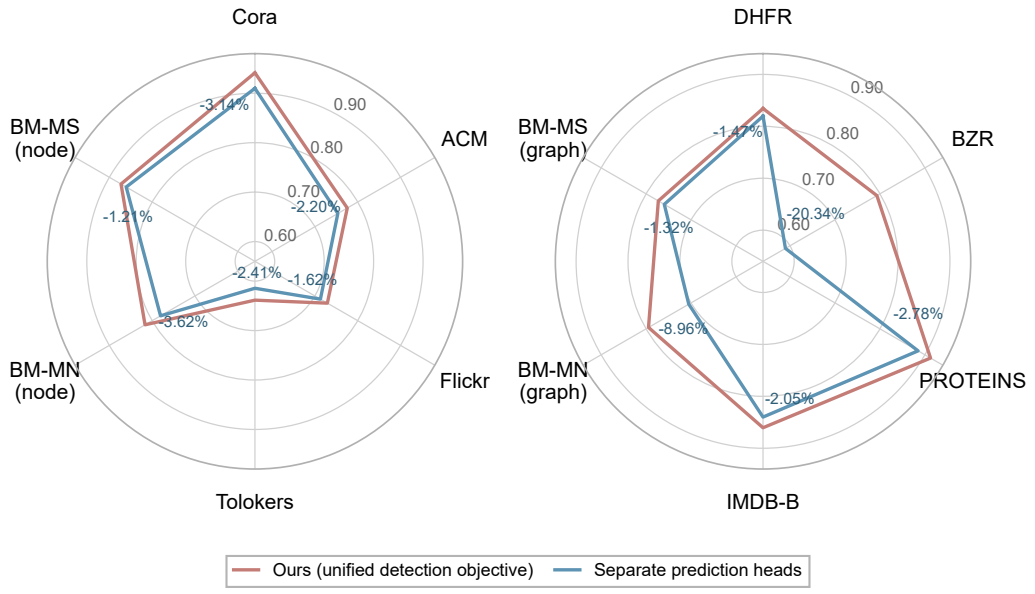


Figure 7: Ablation study on the unified detection objective.

E.1 Additional Main Experiment Results

In this section, we provide a comprehensive comparison of UniTRACER against the baselines in terms of AUROC and AUPRC. Table 3 summarizes the results for node- and edge-level anomaly

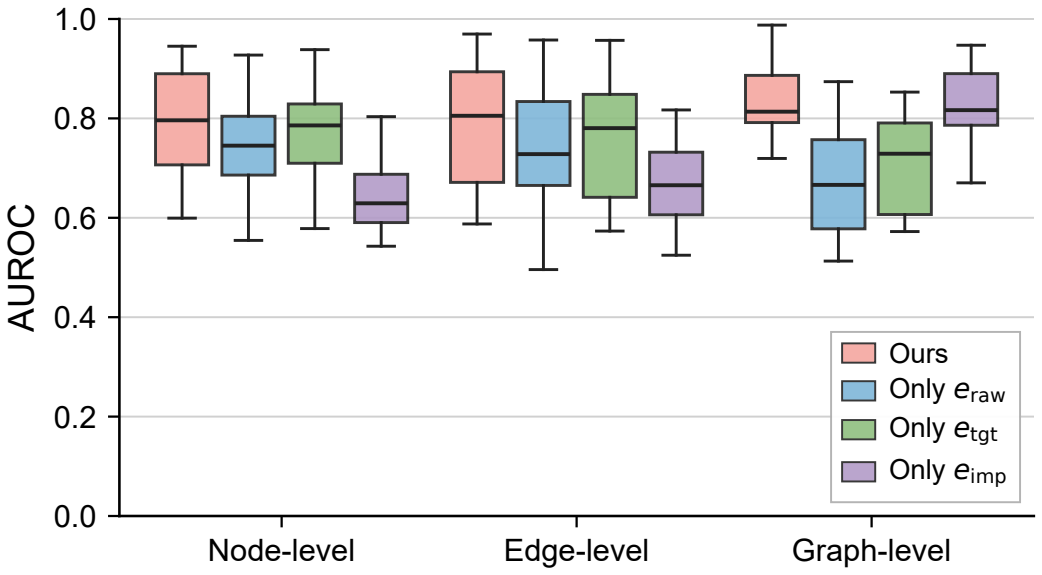


Figure 8: Additional Ablation on Evidence Dimensions.

detection, while Table 4 reports the results on nine GLAD datasets. Overall, UniTRACER achieves strong and consistent anomaly detection performance across both AUROC and AUPRC.

E.2 Additional Ablation on Unified Detection Objective

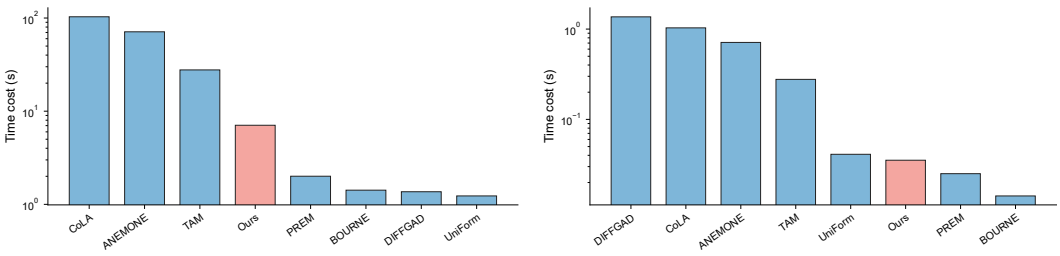
In this section, we further examine the effectiveness of the unified detection objective in UniTRACER. To this end, we construct an ablation variant that replaces the unified objective with separate prediction heads, where the shared autoencoder is replaced by multiple autoencoders for each level of task, which are trained independently with the corresponding level of training task. As shown in Figure E.2, the unified detection objective consistently achieves better performance than the variant with separate prediction heads across all datasets. Despite being trained unsupervised, the unified training objective enables UniTRACER to better capture and integrate anomaly-related knowledge across different structural levels, thereby outperforming other fragmented design that hinder end-to-end knowledge integration.

E.3 Additional Ablation Study on Evidence Dimension

As illustrated in Figure 8, we further evaluate the contribution of each evidence dimension by retaining only one type of evidence at a time. Overall, using a single evidence dimension leads to inferior performance compared with the full model across different anomaly levels, further demonstrating the complementary nature of the three evidence dimensions. Among the individual evidences, $\mathrm{e}_{\mathrm{raw}}$ and $\mathrm{e}_{\mathrm{tgt}}$ provide relatively stronger performance for node- and edge-level anomaly detection, whereas e_{imp} performs particularly well at the graph level. This observation is consistent with the fact that graph-level anomalies are often characterized by a small number of discriminative local structures, making importance-aware evidence especially useful for identifying anomalous graphs. Nevertheless, none of the individual evidence dimensions consistently performs best across all settings, highlighting the benefit of jointly modeling multiple evidence dimensions to capture complementary aspects of anomaly semantics.

E.4 Additional Efficiency Experiment Results

To further examine the efficiency of UniTRACER, we compare the total and per-epoch training time on the Cora dataset. As shown in Figure 9, UniTRACER is substantially faster than computationally intensive baselines such as CoLA, ANEMONE, and TAM in both total and per-epoch training time, while per-epoch running time is comparable to UniForm. Considering the performance gain, these



(a) Total training time. (b) Training time per epoch.

Figure 9: Training time on Cora.

results demonstrate that UniTRACER achieves a favorable trade-off between detection effectiveness and computational efficiency.

E.5 Additional Visualization on Representation Divergence

To investigate how the multi-level tasks are unified by the task-context join modeling and evidence reconstruction module, we conduct additional visualization on inter-granularity representation divergence using Euclidean distance normalized by intra-granularity distance. As shown in Figure 10, UniTRACER consistently reduces the semantic divergence between normal and anomalous classes across different granularities. Overall, the node- and edge-level semantics remain relatively similar in both cases, as anomalous/normal nodes and edges are usually closely coupled in attributed graphs. In contrast, graph-level semantics exhibit substantially larger divergence. The raw node attributes produce the largest divergence. After the encoder, such divergence remains large since the module is designed to preserve granularity-specific information within a unified task format. With the evidence construction, however, the inter-granularity semantic divergence is significantly reduced. This demonstrates that the shared detection evidence effectively transforms the learned target-context information into a shared evidence space with consistent interpretation across detection levels.

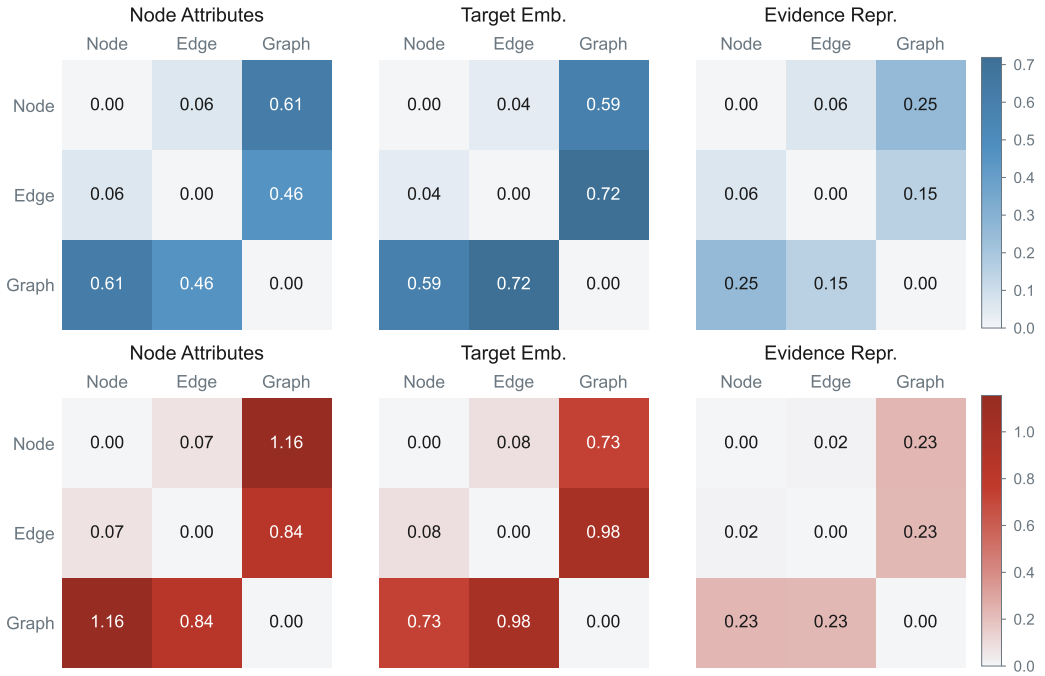


Figure 10: Semantic divergence between levels on BM-MN.